Research Statement

In order to address the needs of twenty-first century scientific computing, the US Department of Energy, machine learning models must be able to "expose biases, and to quantify uncertainties, errors, and precision." [ASCAC, 2020] However, uncertainty quantification and model interrogation can can become quite expensive in high-dimensional machine learning models [Baker et al., 2019, Section 2.4]. Consider, as motivating examples, the following ubiquitous data science tasks.

- Cross validation (CV) is a fundamental tool in machine learning to evaluate model predictive performance and tune hyperparameters, but requires fitting a model multiple times with different data subsets left out.
- Prior specification encodes key assumptions in Bayesian statistics, a paradigm for coherently solving statistical inverse problems. But Bayesian inference can be sensitive to prior specification, particularly in high-dimensional models, and solving the inverse problem for multiple plausible prior choices can be computationally prohibitive.
- Uncertainty propagation, i.e., allowing the inferential uncertainty in one modeling quantity to inform the inferential uncertainty in another, is a key advantage of Bayesian statistics. However, the classical tool for Bayesian estimation, Markov Chain Monte Carlo (MCMC), requires evaluating a statistical model many times, and so can be computationally expensive.

These three central data science problems share the common property that they are computationally demanding due to requiring the evaluation or estimation of a statistical model multiple times: once for each cross validation sample, once for each prior specification, or once for each draw of an MCMC chain.

In my research, I circumvent the computational difficulties of these and other core tasks in data science by using *Taylor series approximations* to extrapolate to nearby couterfactual model inputs (e.g., a new dataset with some points left out, or a new prior specification). By evaluating the derivatives necessary to form the Taylor series at a *single model estimate*, I avoid the re-estimation or re-evaluation that makes the above procedures computationally prohibitive. In exchange, evaluating the necessary derivatives typically requires solving a large but sparse linear system, a tradeoff that can be quite favorable in practice, providing good accuracy orders of magnitude faster than the corresponding classical procedures.

The idea of extrapolating with Taylor series expansions is a venerable one, though the breadth of its potential for contemporary data science problems is arguably underappreciated. My work advances existing research by providing practical implementations of classical methods, particularly using automatic differentiation [Baydin et al., 2017], by updating classical theory to apply in finite sample and under more realistic conditions, and by drawing connections between superficially disparate applications of sensitivity analysis. For the remainder of the essay, I will discuss in more detail my contributions to the above three data science tasks and more, both in practice and in theory.

Approximate cross validation. The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) and the bootstrap can be thought of as examples of this technique.

In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster than repeated re-fitting by an order of magnitude. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the infinitesimal jackknife on a wide variety of problems. As a corollary, I state mild regularity conditions under which the approximation consistently estimates true leave-k-out cross-validation for any fixed k. I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics [Luan and Li, 2003, Shoemaker et al., 2015].

Prior sensitivity for Markov Chain Monte Carlo. MCMC is arguably the most commonly used computational tool to estimate Bayesian posteriors, which is made still easier by modern black-box MCMC tools such as `Stan` [Carpenter et al., 2017, Stan Development Team, 2020]. However, a single run of MCMC typically remains time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models.

My software package, `rstansensitivity`, [Giordano, 2018, Giordano et al., 2018b], takes advantage of the automatic differentiation capacities of `Stan` [Carpenter et al., 2015] together with a classical result from Bayesian robustness [Gustafson, 1996, Basu et al., 1996, Giordano et al., 2018a] to provide automatic hyperparameter sensitivity for generic `Stan` models from only a single MCMC run. I demonstrate the speed and utility of the package in detecting excess prior sensitivity, particularly in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

Prior sensitivity for discrete Bayesian nonparametrics. A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset. A Bayesian nonparametric (BNP) model addresses this question by placing a generative process on cluster assignment, making the number of distinct clusters present amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor a greater or lesser number of distinct clusters.

In [Giordano et al., 2018c], I derive prior sensitivity measures for a truncated variational Bayes approximation using ideas from [Gustafson, 1996, Giordano et al., 2018a]. Unlike previous work on local Bayesian sensitivity for BNP [Basu, 2000], I pay special attention to the ability of the sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure

of robustness *per se*. In work currently in progress, my co-author and I apply the approximation from [Giordano et al., 2018c] to an unsupervised clustering problem on a human genome dataset [Huang et al., 2011, Raj et al., 2014], demonstrating that the approximate is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity.

Uncertainty propagation in mean-field variational Bayes. Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale scientific data sets (e.g., Raj et al. [2014], Kucukelbir et al. [2017], Regier et al. [2019]). However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances [Wang and Titterington, 2004, Turner and Sahani, 2011] due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal model perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. The estimates for MFVB posterior covariances rely on a result from the classical Bayesian robustness literature that relates derivatives of posterior expectations to posterior covariances and includes the Laplace approximation as a special case. In the experiments, I demonstrate that my methods are simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC, including models from ecology [Kéry and Schaub, 2011], the social sciences [Gelman and Hill, 2006], and on a massive internet advertising dataset [Criteo Labs, 2014].

Data ablation. In Giordano et al. [2020], I propose a method to assess the sensitivity of statistical analyses to the removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so I provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. I provide explicit finite-sample error bounds on my approximation for linear and instrumental variables regressions. I demonstrate that non-robustness to data ablation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not a product of misspecification.

The approximation is automatically computable and works for common estimators (including OLS, IV, GMM, MLE, and variational Bayes), and I provide an easy-to-use `R` package to compute the approximation [Giordano, 2020]. Several empirical applications based on published econometric analyses [Angelucci and De Giorgi, 2009, Finkelstein et al., 2012, Meager, 2019] show that even 2-parameter linear regression analyses of randomized trials can be highly sensitive. While I find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even

when standard errors are small.

Frequentist variability of Bayesian posteriors. The frequentist (i.e., sampling) variance of Bayesian posterior expectations differs in general from the posterior variance even for large datasets, particularly when the model is misspecified or contains many latent variables [Kleijn and van der Vaart, 2006]. Knowing the frequentist variance of a posterior expectation can be useful even to a committed Bayesian, particularly when the data is known to arise from random sampling and there is a possibility of model misspecification [Waddell et al., 2002]. However, the principal existing approach for computing the frequentist variability from MCMC procedures is the bootstrap, which can be extremely computationally intensive due to the need to run hundreds of extra MCMC procedures [Huggins and Miller, 2019].

In [Giordano and Broderick, 2020a,b], I propose an efficient alternative to bootstrapping an MCMC procedure which is based on the influence function from sensitivity analysis. Using results from [Giordano et al., 2018a, 2019b], I show that the influence function for posterior expectations can be easily computed from the posterior samples of a single MCMC procedure and consistently estimates the bootstrap variance. I demonstrate the accuracy and computational benefits of the influence function variance estimates on array of experiments including an election forecasting model [Gelman and Heidemanns, 2020], the Cormack-Jolly-Seber model from ecology [Kéry and Schaub, 2011], and a large collection of models and datasets from the social sciences [Gelman and Hill, 2006].

Selected Future work

My research is driven by the needs of my scientific collaborators, and so my future work will be determined to a large part by my colleagues. Here, I will discuss a few directions that I find promising and interesting, and which I believe could be applicable to a diverse set of problems.

The higher-order infinitesimal jackknife for the bootstrap. In the preprint Giordano et al. [2019a], I extend Giordano et al. [2019b] to higher-order Taylor series approximations, providing a family of estimators which I collectively call the higher-order infinitesimal jackknife (HOIJ). In addition to providing higher-quality approximations to CV and extending the results to k-fold CV, the higher-order approach promises to provide a scalable alternative to the bootstrap, a procedure that estimates frequentist variability by repeatedly re-evaluating a model at datasets drawn with replacement from the observed data. The bootstrap is known to enjoy higher-order accuracy in certain circumstances Hall [2013], and the HOIJ can approach the bootstrap at a rate faster than the bootstrap approaches the truth. The HOIJ thus promises to make bootstrap inference available to models which are differentiable but too expensive to re-evaluate (e.g. simulation-based models [Gourieroux and Monfort, 1993]), but also to allow efficient bootstrap-after-bootstrap procedures which that are

currently out of reach for all but the simplest statistics [Efron and Tibshirani, 1994].

Sensitivity for non-differentiable preprocessing. Analyses in genomics often begin with a pre-processing step in observation units are clustered together according to ad-hoc measures of similarity across a large number of feature vectors [Xu and Su, 2015, Stuart et al., 2019]. Quickly assessing the sensitivity of such procedures to the inclusion or exclusion of individual features would allow the researcher to identify high-leverage observations and avoid imposing structure via arbitrary modeling assumptions. However, ordinary sensitivity analysis cannot be applied directly to the clustering step, which is typically non-differentiable.

With a colleague from biology, I am currently investigating a technique that would use importance sampling to compute the sensitivity of such a non-differentiable pre-processing step. We first form a probabilistic relaxation only of the similarity measures, and then run the non-differentiable clustering for an ensemble of Monte Carlo samples of the similarities. Removing a feature would change the probabilities of the similarity measure draws. By differentiating the importance sampling estimate of the effect of the changing probabilities, we can form black-box sensitivity measures with little extra computation other than clustering the original similarity ensemble. Ideally, this sensitivity analysis would allow for quick exploration of the high-dimensional space of feature inclusion, similar to our work in Giordano et al. [2020].

Scaling sensitivity measures. Sensitivity analysis typically avoids the expense of re-fitting a model, but incurs the expense of solving one or several linear systems. Thus, extending the benefits of the sensitivity analysis to increasingly large scientific problems requires developing methods to efficiently solve correspondingly large linear systems. Stochastic second-order methods are currently an active research topic in optimization [Agarwal et al., 2017, Berahas et al., 2020], and methods developed therein should apply directly to sensitivity analysis. I believe these methods would be most fruitfully explored in the context of a particular application, e.g. the production of astronomical catalogs, which I will now discuss.

Partitioned Bayesian inference. The ideas of [Giordano et al., 2018a] can be naturally extended to approximately propagate uncertainty among separately estimated components of an inference problem. For example, astronomical catalogs are customarily produced with MFVB-like algorithms [Lang et al., 2016, Regier et al., 2019], which take inputs such as the sky background and optical point spread function as fixed inputs, though these quantities are themselves inferred with uncertainty. Viewing all the separate inference procedures as a sequential quasi-MFVB objective, one could directly apply the techniques of LRVB to propagate the uncertainty from the modeling inputs to the astronomical catalog's uncertainty.

References

Agarwal, N., Bullins, B., and Hazan, E. (2017). Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187.

Angelucci, M. and De Giorgi, G. (2009). Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508.

ASCAC (2020). US Department of Energy Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee on AI/ML Data-intensive Science and High-Performance Computing, Final draft of report to the committee. Technical report, USDOE Office of Science (SC), Washington, DC (United States).

Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Wilcox, K., and Lee, S. (2019). Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC), Washington, DC (United States).

Basu, S. (2000). Bayesian robustness and Bayesian nonparametrics. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–153.

Berahas, A., Bollapragada, R., and Nocedal, J. (2020). An investigation of Newton-sketch and subsampled Newton methods. *Optimization Methods and Software*, pages 1–20.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carpenter, B., Hoffman, M., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164*.

Criteo Labs (2014). Criteo conversion logs dataset. Downloaded on July 27th, 2017.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. CRC press.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the us elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Giordano, R. (2018). StanSensitivity: automated hyperparameter sensitivity for Stan models. GitHub repository `https://github.com/rgiordan/StanSensitivity`.

Giordano, R. (2020). zaminfluence. GitHub repository `https://github.com/rgiordan/zaminfluence`.

Giordano, R. and Broderick, T. (2020a). The Bayesian infinitesimal jackknife for variance. *In preparation.*

Giordano, R. and Broderick, T. (2020b). Effortless frequentist covariances of posterior expectations in stan. Presentation at Stancon 2020 https://docs.google.com/presentation/d/17Gr8Mqi1yVWliC3SDuHwigCKj1s7RLXgsBhVCEyT5i4/edit?usp=

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Giordano, R., Broderick, T., and Jordan, M. I. (2018b). Automatic robustness measures in stan. Presentation at Stancon 2018 https://docs.google.com/presentation/d/1bxeFy-awELpGlDXNcdJ3rlrKAqNCZZggD8Xmz9Omg0/edit?usp=sharing.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2018c). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *arXiv preprint arXiv:1810.06587*.

Giordano, R., Meager, R., and Broderick, T. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *In preparation.*

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.

Gourieroux, C. and Monfort, A. (1993). Simulation-based inference: A survey with special reference to panel data models. *Journal of Econometrics*, 59(1-2):5–33.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Huang, L., Jakobsson, M., Pemberton, T., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J., Tishkoff, S., and Rosenberg, N. (2011). Haplotype variation and genotype imputation in African populations. *Genetic epidemiology*, 35(8):766–780.

Huggins, J. and Miller, J. (2019). Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Lang, D., Hogg, D., and Mykytyn, D. (2016). The Tractor: Probabilistic astronomical source detection and measurement. *ascl*, pages ascl–1604.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Shoemaker, J. E., Fukuyama, S., Eisfeld, A. J., Zhao, D., Kawakami, E., Sakabe, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). An ultrasensitive mechanism regulates influenza virus-induced inflammation. *PLoS Pathogens*, 11(6):1–25.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*.

Waddell, P., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ml and Bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wang, B. and Titterington, M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, pages 373–380.

Xu, C. and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980.